intellibooks

The New RAG

Traditional RAG

- Splits documents into chunks
- Uses embeddings and vector infrastructure
- Harder to explain retrieved answers
- Adds tuning and infra overhead

1. Document
2. Data Chunking
3. Vector Embeddings
4. Vector Database
5. Similarity Search
6. Ranked Chunks
7. LLM Output

Vectorless RAG

- Reads the structure of the document first
- Navigates directly to the right section
- Preserves context instead of breaking it apart
- Easier to explain, validate, and govern

1. Document
2. Structured Index
3. Intent & Query Routing
4. Hierarchical Navigation
5. LLM Reasoning
6. Precise Retrieval
7. LLM Output

Traditional RAG

Best for discovering conceptual relationships within massive, unstructured knowledge bases.
e.g. Customer support wikis, healthcare research.

Vectorless RAG

Better for pinpoint accuracy, real-time data syncing, & deterministic workflows.
e.g. Software codebases, invoice extraction.